Reprinted from JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS Vol. 39, No. 2, August 1972

A Counterexample in Discounted Dynamic Programming*

A. HORDIJK AND H. C. TIJMS

Mathematisch Centrum, 2E Boerhaavestraat 49, Amsterdam (0), The Netherlands

Submitted by Richard Bellman

1. INTRODUCTION

We are concerned with a dynamic system which at times $t = 0, 1,...$ is observed to be in one of a possible number of states. Let I denote the space of all possible states. We assume I to be denumerable. If at time t the system is observed in state i then a decision k must be chosen from a given finite set K_i. Let Y_t and Δ_t, $t = 0, 1,...$, denote the sequences of states and decisions.

If the system is in state i at time t and decision k is chosen, then two things happen:

(i) We incur a known cost w_{ik} and

(ii) $P\{Y_{t+1} = j \mid Y_0, \Delta_0,..., Y_t = i, \Delta_t = k\} = q_{ij}(k)$, where the $q_{ij}(k)$'s are known.

Finally there is specified a discount factor α, $0 < \alpha < 1$, so that a unit of value at time $t = n$ has a value of α^n at time $t = 0$.

A rule R for controlling the system is a set of non-negative functions $D_k(Y_0, \Delta_0,..., Y_t)$, $k \in K_{Y_t}$; $t \geqslant 0$, where in every case $\sum_k D_k(\cdot) = 1$. As part of a controlling rule, $D_k(Y_0, \Delta_0,..., Y_t)$ is the instruction at time t to make decision k with probability $D_k(Y_0, \Delta_0,..., Y_t)$ if the particular history $Y_0, \Delta_0,..., Y_t$ has occurred.

Let C denote the class of all possible rules. Let C^M denote the class of all memoryless rules, i.e., $D_k(Y_0, \Delta_0,..., Y_t = i) = D_{ik}^{(t)}$ independent of the past history except for the present state. A nonrandomized stationary rule is a memoryless rule for which $D_{ik}^{(t)} = D_{ik}$ independent of t, and in addition $D_{ik} = 1$, or 0 for all i, k.

For any rule $R \in C$ and state $i \in I$, let

$$\psi(i, \alpha, R) = \sum_{t=0}^{\infty} \alpha^t \sum_{j,k} w_{jk} P_R(Y_t = j, \Delta_t = k \mid Y_0 = i),$$

* Report BW 7/71 of the Mathematical Centre, Amsterdam.

provided it exists. The quantity $\psi(i, \alpha, R)$ represents the expected total discounted cost when the initial state is i and rule R is used.

We say that a rule $R^* \in C$ is optimal if $\psi(i, \alpha, R^*) \leqslant \psi(i, \alpha, R)$ for all $R \in C$, $i \in I$.

It is known [1, 2] that there exists an optimal nonrandomized stationary rule when the cost function w_{ik} is bounded. We shall show that an optimal rule may not exist if the boundedness condition on $\{w_{ik}\}$ is weakened. The counterexample given in [2] does not show this result, but proves only that an optimal nonrandomized stationary rule may not exist if the cost function w_{ik} is not bounded. In that counterexample the rule R, which makes with probability $1/(2+t)$ decision 2 when in state i_a at time t, is optimal, since $\psi(i_a, \alpha, R) = -\infty$ for all states i_a.

We shall now give our counterexample.

2. Counterexample

$$I = \{1, 1', 2, 2', \ldots\}, \qquad K_{i'} = \{1\}, \qquad K_i = \{1, 2\}, \qquad i \geqslant 1,$$

$$q_{i'i'}(1) = q_{i,i+1}(1) = 1, \qquad q_{ii'}(2) = 1, \qquad i \geqslant 1,$$

$$w_{i'1} = w_{i1} = 0, \qquad w_{i2} = -\left(1 - \frac{1}{i}\right)\alpha^{-i}, \qquad i \geqslant 1.$$

Clearly, $\psi(i', \alpha, R) = 0$ for all $i \geqslant 1$, $R \in C$. Next we shall prove

$$\psi(i, \alpha, R) > -\alpha^{-i} \qquad \text{for all } i \geqslant 1, R \in C, \tag{1}$$

and

$$\inf_{R \in C} \psi(i, \alpha, R) = -\alpha^{-i} \qquad \text{for all } i \geqslant 1. \tag{2}$$

Since the proof of Theorem 2 in [3] holds also for a denumerable state space, for every $i_0 \in I$ and $R_0 \in C$ there exists a $R \in C^M$ such that

$$P_R(Y_t = i, \Delta_t = k \mid Y_0 = i_0) = P_{R_0}(Y_t = i, \Delta_t = k \mid Y_0 = i_0)$$

for every i, k and t. Hence it suffices to prove (1) for $R \in C^M$.

Let rule $R \in C^M$ and state $i \in I$ be fixed. Denote by $P_i(t)$ the probability that R makes decision 1 when in state $i + t$ at time t. If $P_i(t) = 1$ for all $t \geqslant 0$, then $\psi(i, \alpha, R) = 0 > -\alpha^{-i}$. Suppose now $P_i(t) < 1$ for at least one t. We have

$$\psi(i, \alpha, R) = \sum_{t=0}^{\infty} -\alpha^t\{1 - P_i(t)\} \prod_{k=0}^{t-1} P_i(k) \left(1 - \frac{1}{i+t}\right) \alpha^{-(i+t)}.$$

Using the identity

$$\sum_{t=0}^{\infty} \{1 - P_i(t)\} \prod_{k=0}^{t-1} P_i(k) = 1 - \prod_{t=0}^{\infty} P_i(t),$$

we obtain

$$\psi(i, \alpha, R) > - \alpha^{-i} \sum_{t=0}^{\infty} \{1 - P_i(t)\} \prod_{k=0}^{t-1} P_i(k) \geqslant - \alpha^{-i}.$$

We have now proved relation (1).

If R_n denotes the rule: Make always decision 1 in the states 1,..., $n - 1$, and make always decision 2 in the states n, $n + 1$,..., then

$$\psi(i, \alpha, R_n) = - \alpha^{n-i} \left(1 - \frac{1}{n}\right) \alpha^{-n} = - \alpha^{-i} \left(1 - \frac{1}{n}\right), \qquad n \geqslant i, i \geqslant 1.$$

This relation together with (1) proves (2). By (1) and (2), no optimal rule exists.

References

1. D. Blackwell, Discounted dynamic programming, *Ann. Math. Statist.* **36** (1965), 226–235.
2. C. Derman, Markovian sequential control processes—Denumerable state space, *J. Math. Anal. Appl.* **10** (1965), 295–302.
3. C. Derman and R. E. Strauch, A note on memoryless rules for controlling sequential control processes, *Ann. Math. Statist.* **37** (1966), 276–278.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.